Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution

Abstract

Despite their groundbreaking performance for many generative modeling tasks, diffusion models have fallen short on discrete data domains such as natural language. Crucially, standard diffusion models rely on the well-established theory of score matching, but efforts to generalize this to discrete structures have not yielded the same empirical gains. In this work, we bridge this gap by proposing score entropy, a novel loss that naturally extends score matching to discrete spaces, integrates seamlessly to build discrete diffusion models, and significantly boosts performance. Experimentally, we test our Score Entropy Discrete Diffusion models (SEDD) on standard language modeling tasks. For comparable model sizes, SEDD beats existing language diffusion paradigms (reducing perplexity by 25-75%) and is competitive with autoregressive models, in particular outperforming GPT-2. Furthermore, compared to autoregressive mdoels, SEDD generates faithful text without requiring distribution annealing techniques like temperature scaling (around 6-$8\times$ better generative perplexity than un-annealed GPT-2), can trade compute and quality (similar quality with $32\times$ fewer network evaluations), and enables controllable infilling (matching nucleus sampling quality while enabling other strategies besides left to right prompting).